Optical Engineering

SPIEDigitalLibrary.org/oe

Object of interest extraction in low-frame-rate image sequences and application to mobile mapping systems

Peng Li
Cheng Wang

Object of interest extraction in low-frame-rate image sequences and application to mobile mapping systems

Peng Li
National University of Defense Technology
School of Electronic Science and Engineering
Changsha 410073, China
and
Xiamen University
School of Information Science and Technology
Xiamen 361005, China

Cheng Wang
National University of Defense Technology
School of Electronic Science and Engineering
Changsha 410073, China
E-mail: cwang@xmu.edu.cn

Abstract. Here, we present a novel object of interest (OOI) extraction framework designed for low-frame-rate (LFR) image sequences, typically from mobile mapping systems (MMS). The proposed method integrates tracking and segmentation in a unified framework. We propose a novel object-shaped kernel-based scale-invariant mean shift algorithm to track the OOI through the LFR sequences and keep the temporal consistency. Then the well-known GrabCut approach for static image segmentation is generalized to the LFR sequences. We analyze the imaging geometry of the OOI in LFR sequences collected by the MMS and design a Kalman filter module to assist the proposed tracker. Extensive experimental results on real LFR sequences collected by VISAT™ MMS demonstrate that the proposed approach is robust to the challenges such as low frame rate, fast scaling, and large inter-frame displacement of the OOI.
 [DOI: 10.1117/1.OE.51.6.067201]

Subject terms: low-frame-rate; object of interest; extraction; mean shift; segmentation; tracking; graph cuts; mobile mapping system.

Paper 111572 received Jan. 3, 2012; revised manuscript received Mar. 22, 2012; accepted for publication Apr. 10, 2012; published online Jun. 5, 2012.

1 Introduction

In video processing, object of interest (OOI) extraction solve the problem to interpret image sequence content and extract semantically meaningful objects cared by the users through image sequences[1] and possesses broad application:[2–10] video coding, surveillance, entertainment, etc. During past decade, great successes have been achieved on video OOI extraction through commercial video sequences, whose frame rates are around 15 to 30 fps. High inter-frame coherence is the basic assumption of these algorithms.[8,11,12]

Such high-frame-rate (HFR) videos are good at recording high dynamic scenes with moving object. However, many applications focus on the record and analysis of large coverage and high resolution scenes. In these applications, due to limitations of storage and communication bandwidth, low-frame-rate (LFR) image sequences with large size are often a better choice. In these applications, new challenges are raised for OOI extractions. First, low-frame-rate often results in large scaling of the OOI through the sequence and this requests more robust estimation of the scale factor, which is often assumed to be relatively small in successive frames.[13,14] Second, low-frame-rate often causes large inter-frame displacement of the OOI and this violates the tight assumption of the spatiotemporal continuity of the OOI in previous work.[8,11,12] Third, mobile imaging platform are often used, and the static background assumption[2,7] is no longer available. This challenge is compound with the complexity of the background. These challenges are the aim of our work in this paper.

A good example of the LFR application is the mobile mapping system (MMS), which collects geospatial data as well as image sequences of roadside view using cameras and mapping sensors that are mounted on a moving vehicle[15] and have been widely applied for location-based services such as Google Street View.[16] In MMS, the typical frame rate is 0.5 to 1.5 fbs; the frame size is typically larger than 1280×960; the platform speed is 10 to 100 kmh. Common OOIs include road signs, traffic lights, lane lines,[17–20] etc.

In this paper, we propose a novel approach to extract OOI in LFR image sequences. The proposed approach integrates tracking and segmentation in a unified framework. The entire initialization is only drawing a bounding box around the OOI in the keyframe and much simpler compared to those need precise and trivial labeling of the object and background.[3,4,7] The major contributions of this paper are as follows: first, a new object-shaped kernel is proposed for scale-invariant mean shift algorithm to track the OOI through the LFR sequence to keep the temporal consistency; second, the well-known GrabCut[21] approach for static image segmentation is generalized to the LFR sequence to segment the OOI by propagating the color distribution as well as a dual map of labels; third, we analyze the motion model of the OOI in the LFR sequence collected by MMS, where the OOI undergoes large displacement. Then we use a Kalman filter predictor to increase the robustness of the proposed tracker. We apply the proposed OOI extraction algorithm on real-world LFR image sequences collected by VISAT™ MMS.

This rest of the paper is organized as follows: Sec. 2 presents a brief review on the previous work; Sec. 3 is the overview of the presented approach. Section 4 proposes the object-shaped kernel-based scale-invariant mean shift tracking scheme and the Kalman filter prediction module. Section 5 presents the generalized GrabCut object segmentation method for LFR image sequences. Experimental results on LFR image sequences collected by MMS are discussed in Sec. 6. Section 7 concludes the paper.

0091-3286/2012/$25.00 © 2012 SPIE

2 Related Work

Video OOI extraction methods can be categorized into 3-D volume—based and 2-D frame-by-frame-based groups. Graph cuts-based methods are the foundation of our work and the most representative extraction approaches since the seminal work of Boykov.[22,23] So we only review previous graph cuts-based methods that are directly related to this paper. Then, we have a brief review on mean shift-based tracking algorithms.

2.1 *3-D Volume-Based Approaches*

The 3-D spatiotemporal volume (STV)—based approaches[2–4,24] construct a volume of spatiotemporal by stacking consecutive frames along the temporal dimension and extract the object from the volume directly. Wang et al.[4] present an interactive video cutout system, which provides a novel user interface letting user paint clue strokes of the foreground and background to cut out dynamic foreground objects from a video sequence. The system proposed by Li et al.[3] first oversegments each frame into atomic regions using watershed transform and then performs a 3-D-graph cuts to bi-label each atomic region as foreground or background. The system need the user to specify and segment a keyframe every 10 frames. Although the two above systems can achieve acceptable segmentation results, the application is limited for requirement of quite long preprocessing time for oversegmentation and labor-intensive user interface. Thus, methods of this category are not suitable for OOI extraction from LFR sequence.

2.2 *2-D Frame-by-Frame-Based Approaches*

In 2-D single frame-by-frame based methods,[8,25–27] objects are segmented in image/video sequences frame-by-frame and more practical for consideration of feasibility. In general, pixels in a frame are labeled foreground or background according to color, intensity, texture, temporal consistency, motion or a mixture of them.[7,8,12,25,28] Segmentation results of the previous one or two frames provide prior information of motion, shape, color of the OOI, as well as the temporal consistency inter several adjacent frames. Most existing methods try to integrate this information to the energy term of the Graph Cuts and attempt to seek the global optimal segmentation of the object. Then the segmentation results, often in the form of a labeling trimap, would be propagated to the following frames as the prior of the object. Once some pixels are segmented and labeled incorrectly, the error will be accumulated as propagating through the sequence and deteriorate the results. To keep the temporal consistency, tracking is frequently used to keep the temporal consistency[5,8,25] and will be discussed below.

2.3 *Mean Shift-Based Tracking*

The mean shift algorithm is a nonparametric method based on kernel density estimation (KDE) for mode seeking[29] and successfully applied for visual object tracking through image sequence due to its efficiency and simplicity.[14,30–35]

The original mean shift algorithm is also well known for its limitations of background clutter in radically symmetric kernel, incapability to handle scale change, and large displacement of objects.[32,36] To overcome the first limitation, Yilmaz[32,36] propose an asymmetric level set kernel to exclude background pixels. However, such a kernel cannot meet the requirement of monotonically decreasing of the kernel for some irregular-shaped objects. To overcome the second limitation, Collins proposed to compute the bandwidth of the kernel in the scale space.[34] A better alternation simultaneously estimates the position and scale without the complex scale-space computation.[14,37] To track fast-moving object, a predictor operator is often combined with the mean shift tracker.[31,38,39]

3 Overview of the Proposed Approach

The framework of our approach is illustrated in Fig. 1. The OOI is interactively initialized by a bounding box in one keyframe rather than trivial foreground/background segmentation in several keyframes.[3] The exact OOI is segmented by GrabCut algorithm.[21] Once initialized, the OOI will be extracted automatically.

The OOI extraction scheme integrates tracking and segmentation in a unified framework. Firstly, OOI is tracked by the proposed object-shaped kernel scale-invariant mean shift to keep temporal consistency. The object-shaped kernel increases the robustness and accuracy of mean shift. The adaptive scale selection strategy can handle intensive scaling of the OOI caused by motion of the imaging platform and low frame rate. We further analyze the motion of OOIs in the LFR sequence collected by MMS and propose to use a Kalman filter to augment the robustness of the proposed tracker for tracking fast moving objects. Secondly, the OOI is segmented in the tracked object region by our generalized GrabCut method from the LFR image sequence. The entire tracked region is taken as unlabeled pixels and no foreground is specified, thus we build a dual map rather than trimap as in the graph cuts-based methods.[23] The color distribution is propagated to the next frame instead of the object labels alone as in graph cuts-based algorithms.[7,8,26] Therefore, the original GrabCut approach designed for static

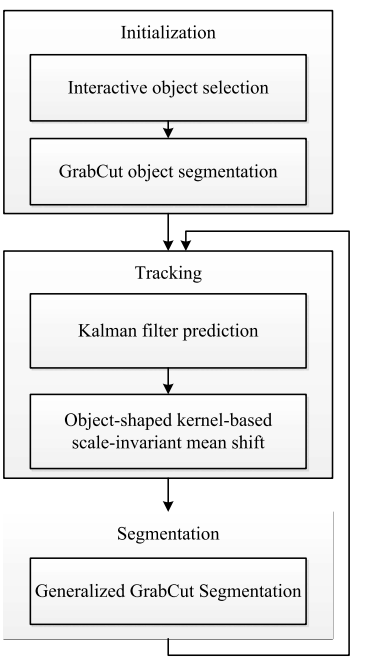


Fig. 1 Diagram of the proposed LFR sequence OOI extraction scheme.

image segmentation is granted the ability to segment object through image sequence. The segmentation result is fed back to the tracking module.

4 Object-Shaped Kernel-Based Scale-Invariant Mean Shift Tracking through LFR Sequence

Original mean shift is prone to fail tracking object with fast changing displacement and size, which are often seen in LFR image sequences due to low frame rate and fast movement of the imaging platform. In this section, we first give a brief review on traditional mean shift tracking methods, and then present our object-shaped kernel-based scale-invariant mean shift scheme. Besides, we investigate the motion of the MMS and propose a Kalman filter predictor to further improve the robustness of the tracker.

4.1 *Brief Review on Mean Shift Object Tracking*

In the traditional mean shift tracking,[35] the object is approximated by an ellipsoidal region and the appearance of the object is represented by its probability density function in a given feature space. A frequently-used appearance model is the m–bin weighted color histogram $\hat{\mathbf{h}}$ because of its low-computational cost and robustness to scaling, rotation, and partial occlusion. Then we have the object model:

$$\begin{cases} \hat{\mathbf{h}} = \{\hat{h}_u\}_{u=1\ldots m} \\ \hat{h}_u = C\sum_{i=1}^{u} k(\|\mathbf{x}_i^*\|^2)\delta[b(\mathbf{x}_i^*) - u] \end{cases}, \tag{1}$$

where $\hat{\mathbf{h}}$ is the object model, $\hat{h}_u$ is the probability of the uth element of $\hat{\mathbf{h}}$, $\{\mathbf{x}_i^*\}_{i=1\ldots n}$ are the normalized pixel locations in the object region centered at the origin point, $k(\mathbf{x})$ is an isotropic kernel with a convex and monotonic decreasing kernel profile, δ is the Kronecker delta function, $b\colon R^2 \to \{1\ldots m\}$ associates to the pixel at location $\mathbf{x}_i^*$ the index $b(\mathbf{x}_i^*)$ of its bin in the quantized feature space. The normalization constant C is derived by imposing the condition $\sum_{u=1}^{m}\hat{q}_u = 1$.

$$C = 1/\sum_{i=1}^{n} k(\|\mathbf{x}_i^*\|^2). \tag{2}$$

Having the weighted color histogram of $\hat{\mathbf{h}}_m$ and $\hat{\mathbf{h}}_c$ generated from the object model and candidate regions, respectively, each pixel $\mathbf{x}_i$ in the candidate region is given a sample weight

$$w(\mathbf{x}_i) = \sum_{u=1}^{m}\sqrt{\frac{\hat{h}_m(u)}{\hat{h}_c(u)}}\delta[b(\mathbf{x}_i) - u]. \tag{3}$$

Let the initial hypothesized position of the object candidate region is centered at $\hat{\mathbf{y}}_{\text{old}}$, the computed new position be $\hat{\mathbf{y}}_{\text{new}}$, the pixels inside the candidate region be $\{\mathbf{x}_i\}$, and then using the sample weight, the mean shift vector $\Delta\mathbf{y} = \hat{\mathbf{y}}_{\text{new}} - \hat{\mathbf{y}}_{\text{old}}$ is calculated as

$$\Delta\mathbf{y} = \frac{\sum_i k'(\|\mathbf{x}_i - \hat{\mathbf{y}}_{\text{old}}\|^2)w(\mathbf{x}_i)(\mathbf{x}_i - \hat{\mathbf{y}}_{\text{old}})}{\sum_i k'(\|\mathbf{x}_i - \hat{\mathbf{y}}_{\text{old}}\|^2)w(\mathbf{x}_i)}, \tag{4}$$

where $k'(\cdot)$ is the derivative of $k(\cdot)$.

4.2 *Object-Shaped Kernel-Based Scale-Invariant Mean Shift Tracking*

4.2.1 *Proposed object-shaped kernel*

In traditional mean shift object tracking,[35] the kernel is commonly chosen as a primitive geometric shape and can't describe the object's shape accurately. The appearance model in Eq. (1) calculated inside the kernel is not accurate for considering the non-object regions residing inside the kernel as a part of the object. To overcome this problem, we propose an object-shaped kernel generated by accurate object mask.

In our work, the object-shaped mask is given by the segmentation results. The centroid of the object is positioned at $\mathbf{x}_c = (x_c, y_c)$, any point $\mathbf{x}_i = (x_i, y_i)$ inside the object region has a distance of $\delta(\alpha_i)$ from the centroid, where α_i is the angle, the point makes with the x-axis. In a similar way, the maximal distance from $\mathbf{x}_c$ in the direction α_i is defined as the radius in α_i

$$r(\alpha_i) = \|\mathbf{x}_b(\alpha_i) - \mathbf{x}_c\|_2, \tag{5}$$

where $\mathbf{x}_b(\alpha_i)$ is the point on the boundary in direction α_i. Using these definitions, the normalized distance $d_{\text{norm}}(x_i)$ in [0,1] from any $\mathbf{x}_i$ to $\mathbf{x}_c$ is defined as

$$d_{\text{norm}}(x_i) = \frac{\delta(\alpha_i)}{r(\alpha_i)} = \frac{\|\mathbf{x}_i - \mathbf{x}_c\|_2}{\|\mathbf{x}_b(\alpha_i) - \mathbf{x}_c\|_2}. \tag{6}$$

In this formalism, the density mode is sought by mean shift in spatial dimension using the Epanechnikov kernel

$$K_E(\mathbf{x}) = \begin{cases} \frac{1}{2}c_d^{-1}(d+2)(1 - \|\mathbf{x}\|^2) & \text{if } \|\mathbf{x}\| = d_{\text{norm}}(\mathbf{x}) \\ 0 & \text{otherwise} \end{cases}, \tag{7}$$

where c_d is the volume of unit d-dimensional sphere. In the mean shift procedure in spatial dimension, we set $d = 2$. Compared to the level set kernel proposed by Yilmaz,[32] where the value of the kernel is the distance from the boundary of the mask and may not guarantee a convex and monotonically decreasing profile in some situation as shown in Fig. 2(b), the object-shaped Epanechnikov kernel generated from (7) has a convex and monotonically decreasing profile. Thus strict convergence is ensured.[29] Moreover, the object-shaped kernel has a very simple derivative and the mean shift vector can be derived by:

$$\Delta\mathbf{y} = \frac{\sum_i w(\mathbf{x}_i)(\mathbf{x}_i - \hat{\mathbf{y}}_{\text{old}})}{\sum_i w(\mathbf{x}_i)}. \tag{8}$$

4.2.2 *Adaptive scale selection*

In the LFR sequence collected by MMS, the OOI often undergoes large scaling and slight orientation change. So the algorithm simultaneously estimates the position and scale of the object as in Yilmaz's work.[32,36]

The scale of an object pixel is represented by a linear transformation of the image coordinates to the scale dimension $s(\mathbf{x}_i)$ of same value as the normalized distance by Eq. (6). There is a scale mean, $\hat{s}$, satisfies the equilibrium of the sum of the pixel-scales on both sides of the scale mean:[32]

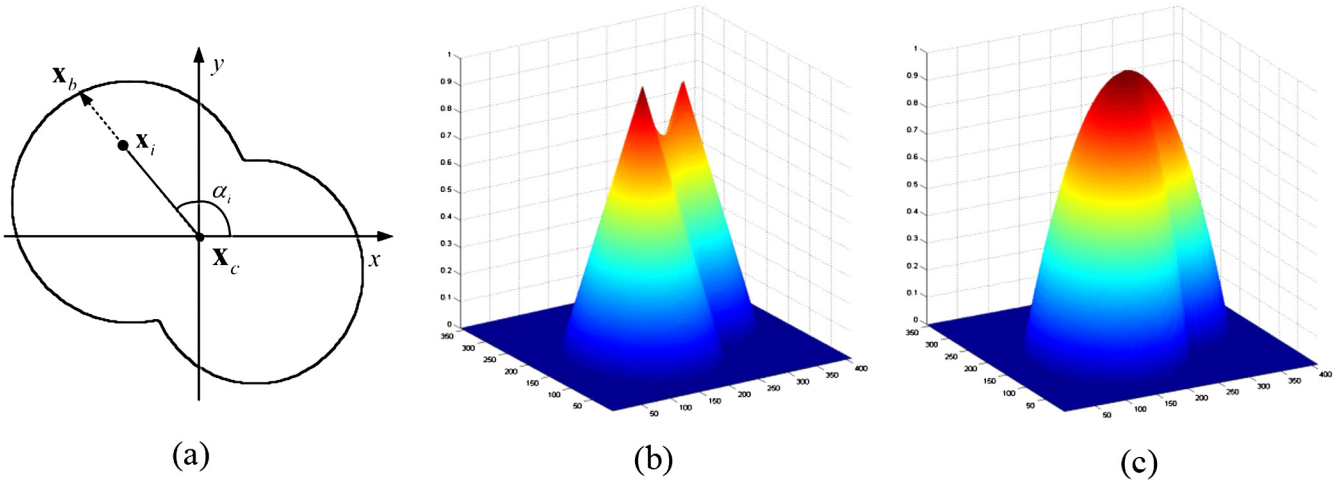


Fig. 2 Object mask and object-shaped kernel. (a) An irregular shape object mask. (b) The level set kernel. (c) The proposed kernel.

$$\int_0^{2\pi}\int_0^{\hat{s}r(\alpha)}\frac{\delta}{r(\alpha)}\mathrm{d}\delta\mathrm{d}\alpha=\int_0^{2\pi}\int_{\hat{s}r(\alpha)}^{r(\alpha)}\frac{\delta}{r(\alpha)}\mathrm{d}\delta\mathrm{d}\alpha. \tag{9}$$

To solve this equation, we obtain the solution of the scale-mean by $\hat{s}=\sqrt{2}^{-1}$. It is a constant for any type of object shape. An update in the scale dimension should take relation between the bandwidth and the scale into consideration. Similar to computing the spatial change, scale change can be estimated by setting the dimensions in the Epanechnikov kernel in Eq. (7) to $d=1$ and replacing the distance d_{norm} with scale s. Let the mean shift iteration provide a scale update of Δs, the new bandwidth $r_{\text{new}}(\alpha)$ will be updated by

$$r_{\text{new}}(\alpha)=(1+\sqrt{2}\Delta s)r(\alpha). \tag{10}$$

4.2.3 *Object tracking in the 3-D spatial-scale space*

To run the mean shift iterations in the joint 3-D spatial-scale space $\mathbf{\Gamma}=(x,y,s)$, a 3-D Epanechnikov kernel consisting of the product of the spatial object-shaped kernel and the kernel for the scale dimension

$$K_{3d}(x,y,s)=K_E(x,y)K_E(s) \tag{11}$$

is defined based on the assumption that the incremental updates to each dimension is independent to each other. Using the 3-D kernel given in (11), the mean shift vector which maximizes the density iteratively is computed by:

$$\Delta\mathbf{\Gamma}=\frac{\sum_i w(\mathbf{\Gamma}_i)(\mathbf{\Gamma}_i-\mathbf{\Gamma}_{\text{old}})}{\sum_i w(\mathbf{\Gamma}_i)}. \tag{12}$$

Benefiting from the proposed object-shaped kernel, the scale s at point $\mathbf{x}_i$ is the same as the normalized distance and calculated once building the object-shaped kernel. Consequently, the efficiency is improved.

Outline of the object-shaped kernel scale-invariant mean shift is given in Fig. 3.

Object-shaped kernel-based scale-invariant mean shift

1 Compute centroid $\mathbf{x}_c$, radius table $\{r(\alpha_i)\}$ and object mask shape kernel
2 Generate object model $\hat{\mathbf{h}}_m$
3 for all frames
4 Loop until convergence
5 Get candidate region with radius table $\{r(\alpha_i)\}$ and candidate model
6 Perform mean shift in $\Gamma=(x,y,s)$ space
7 Update object centroid: $\mathbf{x}_c=\mathbf{x}_c+\Delta\mathbf{x}$
8 Update radius table by new scale: $\{r(\alpha_i)=r(\alpha_i)*(1+\sqrt{2}\Delta s)\}$
9 endLoop
10 endFor

Fig. 3 The proposed object-shaped kernel-based scale-invariant mean shift algorithm.

4.3 *Kalman Filter-Based Prediction*

Kalman filter has often been integrated to mean shift for increasing the robustness of the trackers.[31,38,39] In the proposed Kalman filter module, the prediction result of the Kalman filter is coarse and will be refined by the proposed object-shaped kernel-based scale-invariant mean shift. The two-stage tracking scheme is shown in Fig. 4.

In order to predict the object location by Kalman filter, we analyze motion model of the OOI in the LFR image sequences collected by MMS and construct the dynamic equations. Assuming the surveying vehicle undergoing uniform linear motion and a pin-hole model to the camera, we have the motion model of the object in image plane as shown in Fig. 5.

From Fig. 5, we can see the object moves faster in the image plane and its size becomes bigger as the surveying vehicle approaching. We get the following relationship by pin-hole model.

$$\frac{x_t'}{x_{\text{obj}}}=\frac{f}{z_t},\qquad \frac{y_t'}{y_{\text{obj}}}=\frac{f}{z_t} \tag{13}$$

$$\frac{x_{t+1}'}{x_t'}=\frac{z_t}{z_{t+1}}=\frac{z_t}{z_t-v\Delta t},\qquad \frac{y_{t+1}'}{y_t'}=\frac{z_t}{z_{t+1}}=\frac{z_t}{z_t-v\Delta t} \tag{14}$$

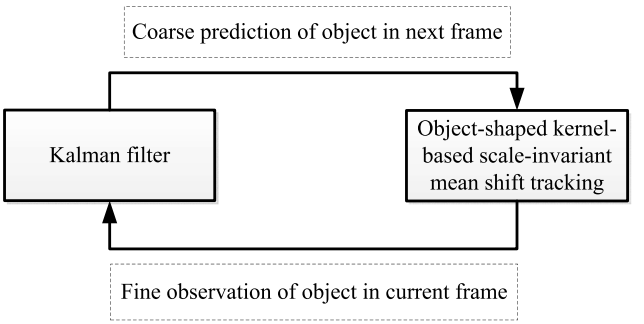


Fig. 4 Kalman filter assisted mean shift.

$$\begin{cases} x'_{t+1} = \frac{z_t}{z_t - v\Delta t} x'_t = \frac{1}{1-(v/z_t)\Delta t} x'_t \\ y'_{t+1} = \frac{z_t}{z_t - v\Delta t} y'_t = \frac{1}{1-(v/z_t)\Delta t} y'_t \end{cases}. \tag{15}$$

We expand Eq. (15) by Taylor expansion to time interval Δt and preserve the first two order items for accuracy and efficiency consideration. We get that x'_{t+1} and y'_{t+1} in the following form

$$\begin{cases} x'_{t+1} = (1 + a_1\Delta t + a_2\Delta t^2)x'_t = x'_t + a_1 x'_t \Delta t + \frac{1}{2} a_2 x'_t \Delta t^2 \\ y'_{t+1} = (1 + a_1\Delta t + a_2\Delta t^2)y'_t = x'_t + a_1 y'_t \Delta t + \frac{1}{2} a_2 y'_t \Delta t^2 \end{cases}. \tag{16}$$

In our case, for a_1 and a_2 are highly related to vehicle velocity v and z_t as, which are difficult to be accurately acquired, we formulate objects in the image plane moving in uniformly accelerated motion and approximate and $a_1 x'_t$ and $a_1 y'_t$ as the velocity in x and y direction, denoted by v'_x and v'_y, respectively. In the same manner, we approximate $a_2 x'_t$ and $a_2 y'_t$ as the acceleration in x and y direction, denoted by a'_x and a'_y. Then we get

$$\begin{cases} x'_{t+1} = x'_t + v'_x \Delta t + \frac{1}{2} a'_x \Delta t^2 \\ y'_{t+1} = x'_t + v'_y \Delta t + \frac{1}{2} a'_y \Delta t^2 \end{cases}. \tag{17}$$

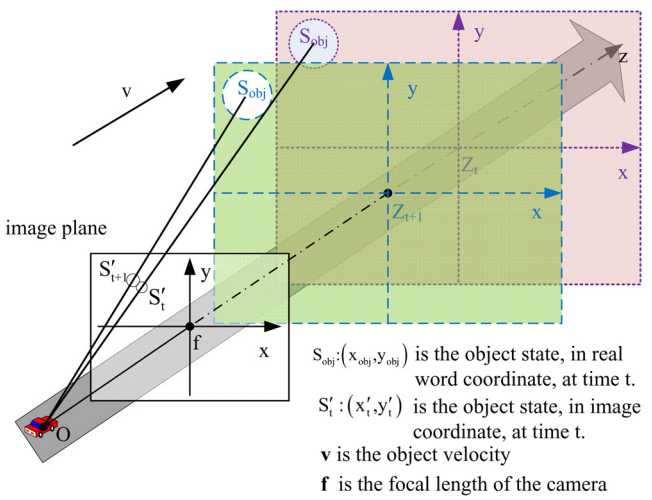


Fig. 5 Motion model of the OOI in LFR sequence collected by MMSs.

In tracking, not only location but also size of the tracked object is estimated. Then the state vector of Kalman filter is $\mathbf{x} = [x, y, v_x, v_y, a_x, a_y, s, v_s]^T$ consists of the coordinates, velocity, acceleration of the object and relative size and velocity of change of relative size. The state-transition Eq. (18) is

$$\begin{bmatrix} x(k+1) \\ y(k+1) \\ v_x(k+1) \\ v_y(k+1) \\ a_x(k+1) \\ a_y(k+1) \\ s(k+1) \\ v_s(k+1) \end{bmatrix} = \begin{bmatrix} 1 & 0 & \Delta t & 0 & \frac{1}{2}\Delta t^2 & 0 & 0 & 0 \\ 0 & 1 & 0 & \Delta t & 0 & \frac{1}{2}\Delta t^2 & 0 & 0 \\ 0 & 0 & 1 & 0 & \Delta t & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & \Delta t & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & \Delta t \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} x(k) \\ y(k) \\ v_x(k) \\ v_y(k) \\ a_x(k) \\ a_y(k) \\ s(k) \\ v_s(k) \end{bmatrix} + N(0, Q). \tag{18}$$

The observations of our method consist of the object's centroid coordinates and the relative size. The observation Eq. (19) is

$$\begin{bmatrix} x(k) \\ y(k) \\ s(k) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x(k) \\ y(k) \\ v_x(k) \\ v_y(k) \\ a_x(k) \\ a_y(k) \\ s(k) \\ v_s(k) \end{bmatrix} + N(0, R). \tag{19}$$

5 Generalized GrabCut for Object Segmentation from LFR Sequences

5.1 *Graph Cuts-Based Video Object Segmentation and its Limitations*

Boykov and Jolly[22] were the first to formulate a simple generative MRF model in discrete domain for the task of binary image segmentation. They addressed the problem of interactive image segmentation based on initial trimap $T = \{T_F, T_B, T_U\}$. The trimap partitions the image into three sets: T_F and T_B comprises of pixels selected by the user as either foreground or background respectively, and T_U is the remaining set of unknown pixels. The segmentation is formulated as estimation of a global minimum problem and solved by the min-cut max-flow algorithm.

(a) (b) (c) (d)

Fig. 6 (a) The yellow brush pot as an instance of OOI, (b) OOI mask, (c) The trimap of the graph cuts-based method, white, gray, and black indicate foreground, the unlabeled and background respectively, and (d) The dual map of the proposed method.

The segmentation results is a mask with pixel labeled as the foreground and the background defined as $S = \{L_F, L_B\}$.

For object segmentation from video sequences, graph cuts-based video object segmentation approaches generally treat the segmentation results in the previous frame as the label of foreground in current frame to guide and accelerate the segmentation,[7,8,25,26] i.e., the trimap in the current ith frame $T(i) = \{T_F(i), T_B(i), T_U(i)\}$ is determined by the segmentation result $S(i-1) = \{L_F(i-1), L_B(i-1)\}$ in the previous $(i-1)$th frame. As shown in Fig. 6(b) and 6(c), pixels around the boarder of the object belong to $T_U(i)$; other pixels in $L_F(i-1)$ and $L_B(i-1)$ belong to $T_F(i)$ and $T_B(i)$, respectively. During the segmentation of the ith frame, only pixels in $T_U(i)$ will be partitioned once while the labels of the other pixels remain unchanged.

The graph cuts-based methods are hindered by the following limitations. Firstly, it requires that the frame rate of the sequences is high enough to ensure the spatiotemporal consistency of the object appearance between consecutive frames.[2,7] Unfortunately, the OOI undergoing large displacement and scaling in LFR image sequences. Direct propagation of the object segmentation results may lay foreground labels on background pixels and vice-versa, and thus cause error segmentation. Secondly, segment error will be propagated through the following sequence and the error will be accumulated. Thirdly, initializations of the graph cuts-based methods need to specify both the foreground and background by trivial sketches.

A more flexible labeling and propagation strategy is needed. An alternative is the GrabCut[21] algorithm, which allows incomplete labeling, i.e. only T_B for the trimap is specified and represents monochrome image model by a Gaussian Mixture Model (GMM) and replaces the one-shot minimum cut estimation algorithm by a more powerful iterative procedure. However, the GrabCut is designed for object segmentation in static image and there is no clear mechanism to use GrabCut to segment OOI in LFR sequences.

5.2 *Generalized GrabCut for Video Object Segmentation*

Motivated by the GrabCut approach for static image segmentation, we propose a generalized GrabCut framework for video object segmentation.

In contrast to the graph cuts-based approaches which build trimap T_U, we construct a dual map $D_{\text{Map}}(i) = \{D_U(i), D_B(i)\}$ for the current ith frame, where D_U and D_B are pixels selected as unlabeled and background, respectively, by inheriting the segmentation result $S(i-1) = \{L_F(i-1), L_B(i-1)\}$ in the previous $(i-1)$th as prior information for current frame. As shown in Fig. 6(c), no foreground pixels are specified. For there is no mandatory labeling of the foreground, the mislabeling error propagation can be avoided by setting the D_U properly larger than original object mask.

Although OOI in LFR sequence changes their locations and sizes dramatically in consecutive frames, their appearance remain relatively invariant. Therefore, the dual map $D_U(i)$ and $D_B(i)$ in frame i inherit the segment results $S(i-1) = \{L_F(i-1), L_B(i-1)\}$ together with their color distribution $G_F(i-1)$ and $G_B(i-1)$ in frame $i-1$, as initialization for segmentation of the next frame, i.e.,

$$G_D(i) = G_D[i|S_D(i-1)], \qquad D \in \{F, B\}. \tag{20}$$

For the ith frame, the image consists of N pixels z_n in RGB color space and the color distribution is modeled by two GMMs, $G_F(i)$ and $G_B(i)$ for the unlabeled pixels and background, respectively. The GMMs are taken to be a full-covariance Gaussian mixture with K components (typically $K = 5$). In order to deal with the GMM tractably, in the optimization framework, an additional vector $\mathbf{k} = \{k_1, \ldots, k_n, \ldots, k_N\}$ is introduced, with $k_n = \{1, \ldots, K\}$, assigning, to each pixel, a unique GMM component, one component either from G_B or the G_F, according to $\alpha_n = 0$ or 1.

The Gibbs energy, which is formulated as the sum of region term R and boundary term B for segmentation now becomes

$$E(\boldsymbol{\alpha}, \mathbf{k}, \boldsymbol{\theta}, \mathbf{z}, i) = R(\boldsymbol{\alpha}, \mathbf{k}, \boldsymbol{\theta}, \mathbf{z}, i) + B(\boldsymbol{\theta}, \mathbf{z}). \tag{21}$$

The data term R is now defined, taking account of the color GMM models, as

$$\begin{aligned} R(\boldsymbol{\alpha}, \mathbf{k}, \boldsymbol{\theta}, \mathbf{z}, i) &= \sum_n D(\alpha_n, k_n, \underline{\theta}, z_n, i) \\ &= \sum_n D(\alpha_n, k_n, \underline{\theta}, z_n | S_{i-1}), \end{aligned} \tag{22}$$

where $D(\alpha_n, k_n, \underline{\theta}, z_n, i) = -\log G(z_n|\alpha_n, k_n, \underline{\theta}, i) - \log \pi(\alpha_n, k_n, i)$, and $G(\cdot)$ is the Gaussian probability distribution, and $\pi(\cdot)$ are mixture weighting coefficients.

The boundary term B is computed using Euclidean distance in color space:

$$B(\alpha, \mathbf{z}) = \gamma \sum_{(m,n)\in C} [\alpha_n \neq \alpha_m] \exp\left(-\frac{\|z_m - z_n\|^2}{\beta}\right). \tag{23}$$

The constant β is taken to be $[2\langle (z_m - z_n)^2 \rangle]^{-1}$ where $\langle \cdot \rangle$ denotes expectation over an image sample.

Having all the parameters defined, the binary segmentation $\hat{\boldsymbol{\alpha}}$ can be formulated as estimation of a global minimum and is solved by max-flow min-cut algorithm.

Generalized GrabCut Approach

0. Read in frame i, the segmentation results $S_D(i-1)$ as well as the distribution of the previous frame $G_D(i-1)$.

1. Assigning GMM components to pixels: for each n in $D_U(i)$.

$$k_n = \arg\min_{k_n} D(\alpha_n, k_n, \theta, z_n, i)$$

2. Estimate segmentation: use max-flow min-cut algorithm to solve:

$$\alpha_n = \arg\min_{\{\alpha_n, n\in T_U\}} E(\boldsymbol{\alpha}, \mathbf{k}, \boldsymbol{\theta}, \mathbf{z}, i)$$

3. Update GMM parameters $\boldsymbol{\theta}$ from the new segmentation $\boldsymbol{\alpha}$.

4. Repeat from step 1, until convergence, e.g. $|\Delta\boldsymbol{\alpha}| < \varepsilon$.

5. Post-processing to eliminate the jitters.

Fig. 7 The generalized GrabCut OOI segmentation scheme.

$$\hat{\boldsymbol{\alpha}} = \underset{\alpha}{\operatorname{argmin}}\, E(\boldsymbol{\alpha}, \mathbf{k}, \boldsymbol{\theta}, \mathbf{z}, i). \tag{24}$$

In the proposed scheme, the energy minimization in Eq. (24) is implemented iteratively until convergence. We adopt morphological post-processing to eliminate the isolated segments and propagate $G_F(i)$ and $G_B(i)$ to next frame on-line. By these means, our generalized GrabCut approach has been granted the ability to segment objects from dynamic image sequences. The outline of the proposed segmentation approach is shown in Fig. 7.

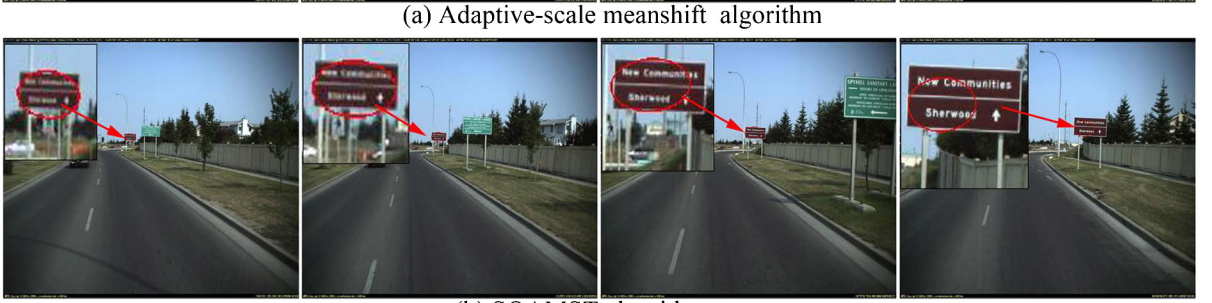
6 Experimental Results

In this section, we evaluate our OOI extraction method with real-world LFR image sequences collected in Calgary, Canada, by the VISAT™ MMS. The size of the frames is 1600×1238, and the frame rate is 1 fps. The MATLAB demos and some test data for this paper are available at http://chwang.xmu.edu.cn/works/demo_ooi_extraction.htm. To re-implement the proposed method, the max-flow/min-cut library is required and available at http://vision.csd.uwo.ca/code/.

6.1 *Tracking Performance Evaluation*

To evaluate the performance of the proposed object-shaped kernel-based scale-invariant mean shift algorithm, we have performed comparisons with the standard adaptive scale mean shift algorithm[35] and a newest scale adaptive mean shift algorithm, SOAMST,[14] which adaptively selects scale by compute the moment of the weight image and is available at http://www4.comp.polyu.edu.hk/~cslzhang/SOAMST.htm. Here, we just present some representative results. In all experiments, the RGB color space was taken as feature space and quantized to $8 \times 8 \times 8$ bins. In these experiments, the OOI are given by a bounding rectangle in only 1 keyframe and is firstly segmented by

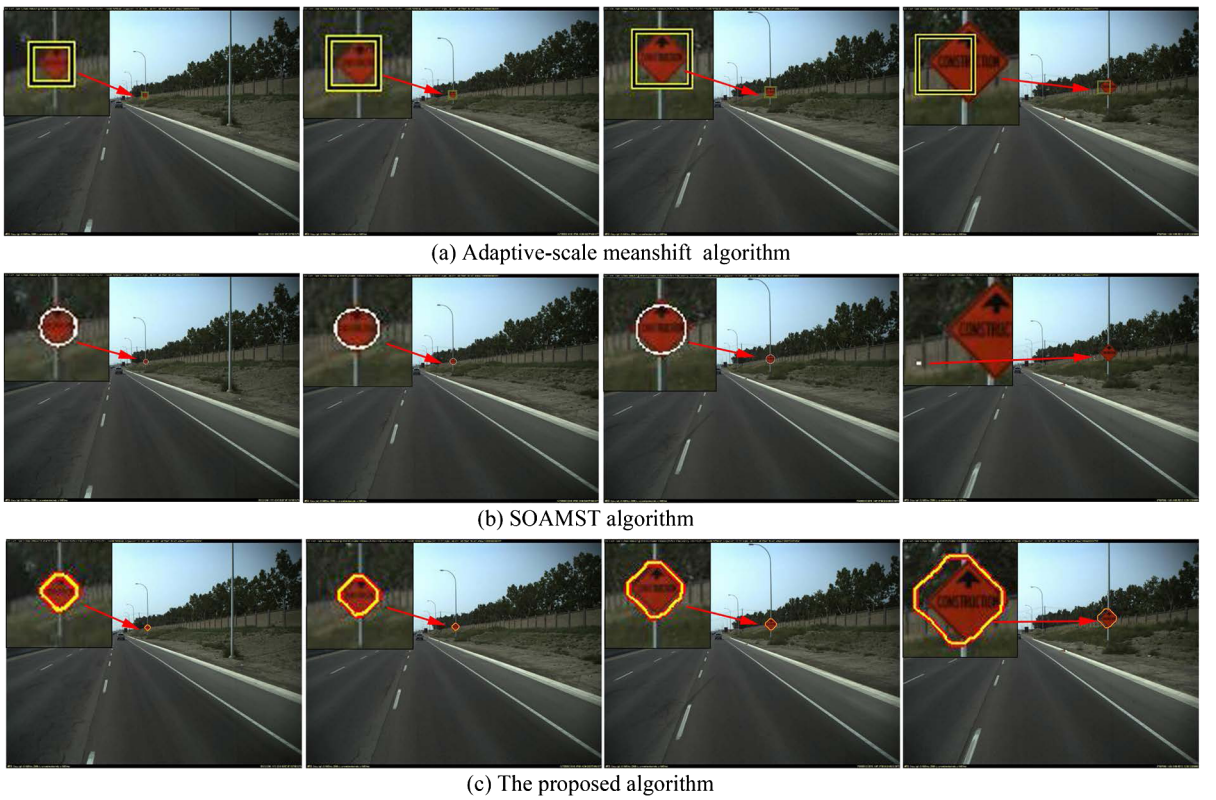
(a) Adaptive-scale meanshift algorithm



(b) SOAMST algorithm

(c) The proposed algorithm

Fig. 8 Tracking results of a brown guideboard. The frames 3, 6, 9, and 12 are displayed.



(a) Adaptive-scale meanshift algorithm

(b) SOAMST algorithm

(c) The proposed algorithm

Fig. 9 Tracking results of an orange guideboard. The frames 10, 15, 18, and 22 are displayed.

GrabCut[21] algorithm to generate the object mask for the proposed kernel-shaped scale-invariant mean shift tracker. The tracking accuracy is evaluated by the distance between the tracked object centroid and the ground truth.

Figures 8 and 9 show the qualitative results of tracking a brown rectangular guideboard and a red diamond guideboard, the tracking details are magnified in the left-top. As shown in Figs. 8 and 9, we can the adaptive-scale mean shift[35] and the SOAMST[14] fails to track the OOI in the last several frames where the objects undergoes large scaling, while our algorithm can track the OOIs until they disappear in the sequences. Quantitative results of the brown and red guideboard sequences are shown in Figs. 10 and 11, respectively. It can be seen that our method outperforms the other two methods significantly.

We also evaluate the performance of the proposed Kalman filter predictor using the above two sequences. We specify the OOI in the first two adjacent frames by a simple bounding box and calculate the initial value of acceleration by solving Eq. (17) given a small initial velocity. As shown

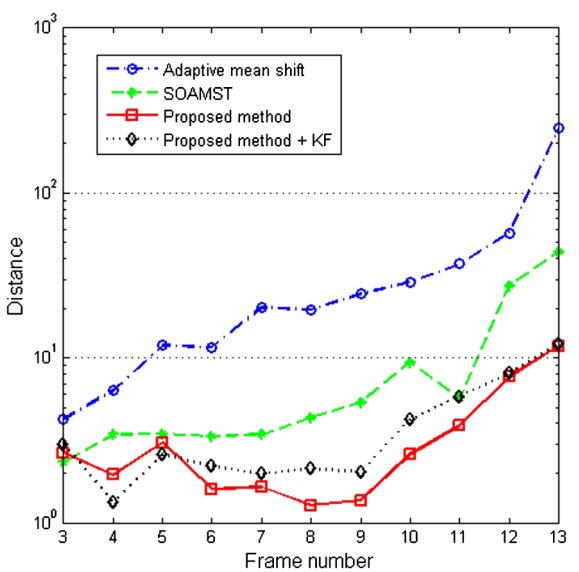


Fig. 10 Tracking accuracy of the brown guideboard.

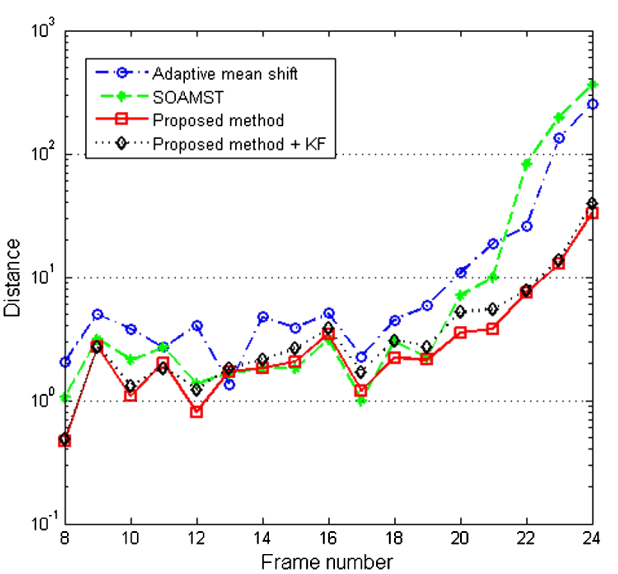


Fig. 11 Tracking accuracy of the red guideboard.

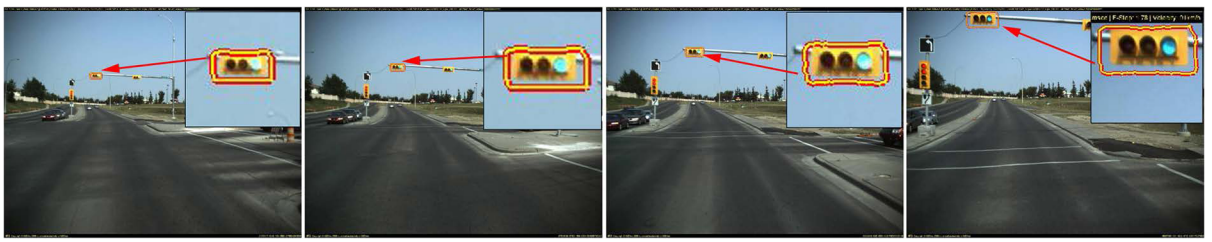


Fig. 12 Tracking results of the proposed Kalman filter assisted tracking scheme. The frames 4, 6, 8, and 10 are displayed.

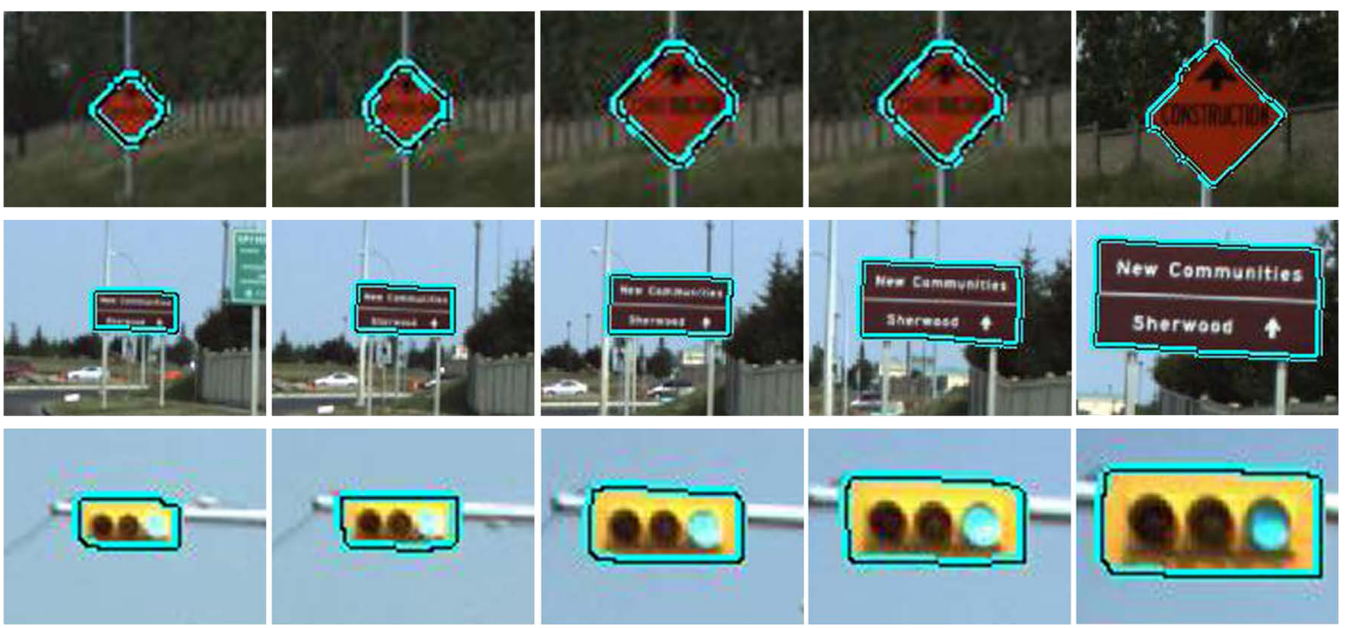


Fig. 13 OOI segmentation results for real-world LFR sequence in MMS.

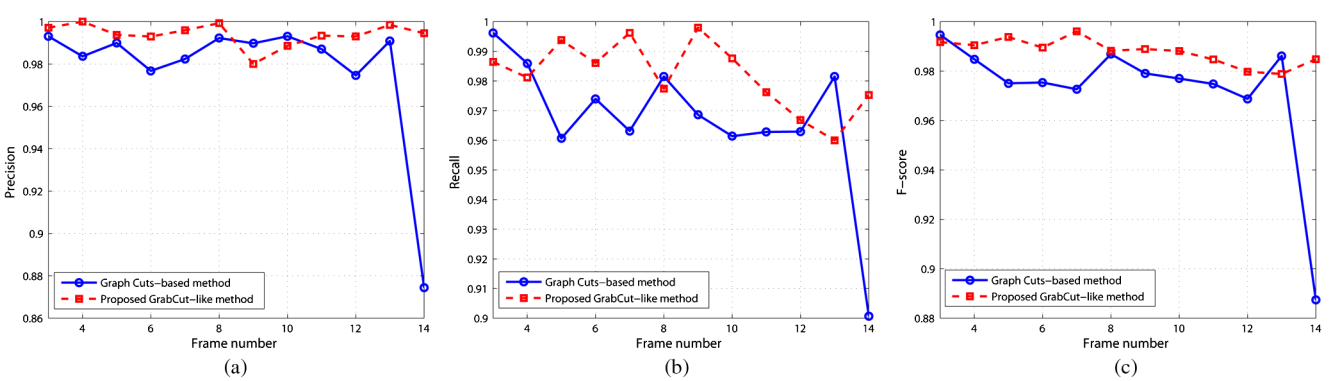


Fig. 14 (a) Precision, (b) recall and (c) *F*-score of the brown guideboard sequence.

in Figs. 10 and 11, the performance is approximately equal to the results achieved without the Kalman filter module because the initial value is not accurately estimated and surveying vehicle is not undergoing ideal uniform linear motion. However, when tracking OOI with large inter-frame displacement and scaling, the proposed object-shaped kernel-based scale-invariant mean shift approach alone fails in some cases. Whereas it is observed that the proposed Kalman module can significantly alleviate this problem since it can give a more reasonable but coarse start point for the fine mean shift tracking procedure. Figure 12 shows the tracking results of a traffic light by the Kalman filter assisted scheme where the object-shaped kernel-based scale-invariant mean shift method fails. In this sequence, the traffic light travels about 500 points through 10 frames.

In practical MMS applications, a more accurate initial state of the Kalman filter module can be estimated from the measurable GIS information, which is beyond the discussion of this paper. So the introduction of Kalman filter is essential and promising.

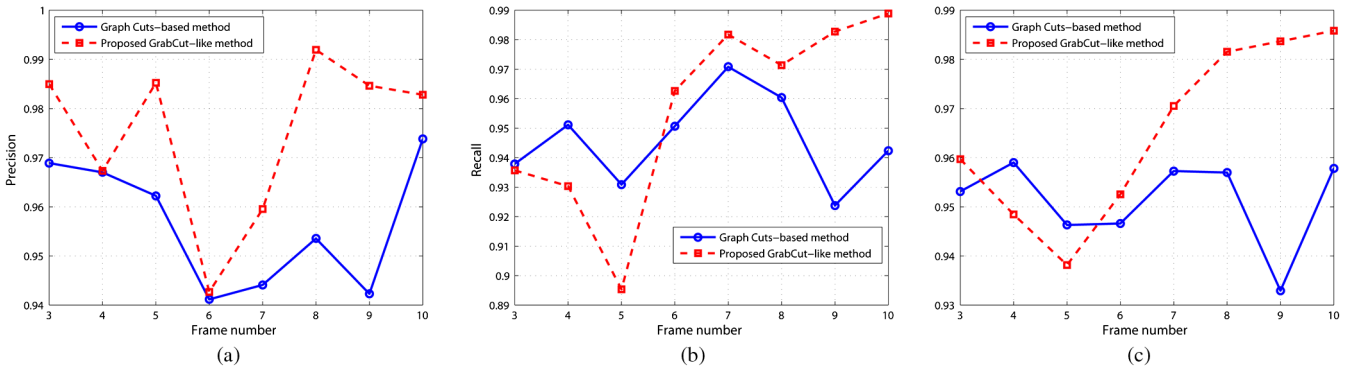


Fig. 15 (a) Precision, (b) recall and (c) *F*-score of the traffic light sequence.

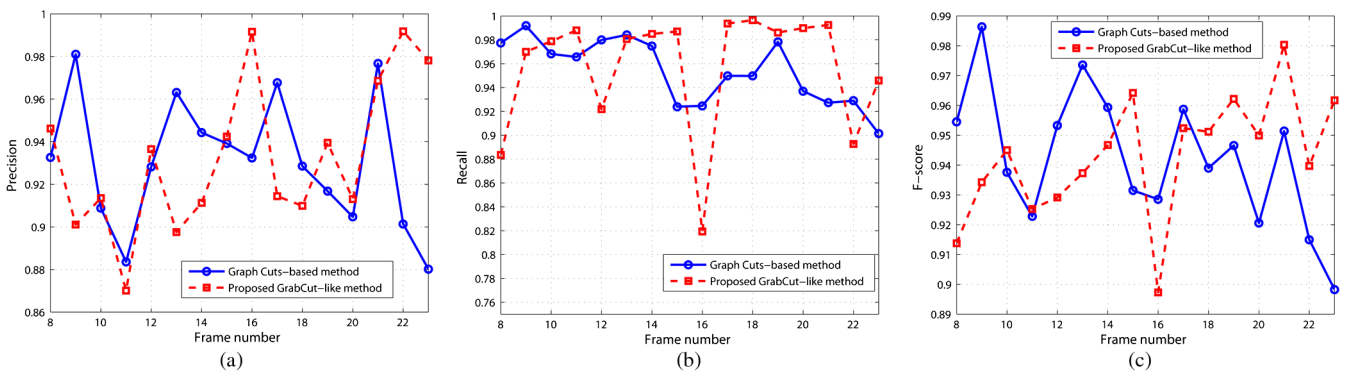


Fig. 16 (a) Precision, (b) recall and (c) *F*-score of the red guideboard sequence.

6.2 *Segmentation Performance Evaluation*

We obtain the candidate region of the OOI in current frame as the result of tracking. Then we segment the object in the candidate region using the proposed generalized GrabCut approach and present the segmentation results of the traffic light, red guide board and brown guide board sequence in Fig. 13.

In order to perform objective evaluation of our approach, we first manually segmented the reference objects (or ground truths) for the test LFR sequences. The segmentation performance is measured by the precision and recall defined in Eqs. (25) and (26) and compared to base-line graph cuts-based methods.[8,22,26] We haven't performed matting as the authors do in GrabCut approach.[21]

$$\text{precision} = \frac{\{\text{true area}\} \cap \{\text{segmented area}\}}{\{\text{segmented area}\}}, \tag{25}$$

$$\text{recall} = \frac{\{\text{true area}\} \cap \{\text{segmented area}\}}{\{\text{true area}\}}. \tag{26}$$

A summary of precision and recall can be plotted in terms of their harmonic mean, which is referred to as the *F*-score:

$$F = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \tag{27}$$

The expectation is to observe the precision, recall and *F*-score values close to 1 for both measures. Figures 14 to 16, show the precision, recall, and *F*-score of the brown guideboard, traffic light, and the red guideboard sequences, respectively. The average precision, recall and *F*-score are shown in Tables 1 to 3.

As shown in Figs. 14 to 16, we can see the proposed generalized GrabCut object segmentation approach can achieve better segmentation in most frames. Especially when the object is moving fast, the graph cuts-based method suffers the wrong-propagated labels and is prone to produce a worse result. From Tables 1 to 3, our method performs better with more stable segmentation than the graph cuts-based methods except for the red guideboard sequence. However, our method reports a better average recall and the two other measures differ little. Moreover, the total interactions involved in our method are much less than in graph cuts-based methods.

Table 1 Average precision, recall and *F*-score of the brown guideboard sequence.

	Graph cuts-based method	Our generalized GrabCut
Average precision	0.9773	0.9939
Average recall	0.9660	0.9821
Average *F*-score	0.9719	0.9879

Table 2 Average precision, recall and *F*-score of the traffic light sequence.

	Graph cuts-based method	Our generalized GrabCut
Average precision	0.9568	0.9749
Average recall	0.9460	0.9561
Average *F*-score	0.9513	0.9651

Table 3 Average precision, recall and *F*-score of the red guideboard sequence.

	Graph cuts-based method	Our generalized GrabCut
Average precision	0.9306	0.9329
Average recall	0.9540	0.9569
Average *F*-score	0.9426	0.9432

7 Conclusions

LFR video applications bring new challenges to OOI extraction, such as large scaling, large inter-frame displacement, and clutter background. In this paper, we propose a novel method to extract OOI in LFR image sequence and delineate its application to the MMS. The method integrates object-shaped kernel-based scale-invariant mean shift tracking and the generalized GrabCut segmentation in a unified framework and needs only one bounding box in the keyframe as initialization. The proposed object-shaped kernel is applied to mean shift tracking with adaptive scale selection. We further analyze the motion model of the OOI in the LFR sequence collected by the MMS and design a Kalman filter module to enhance the robustness of the proposed tracker. Finally, we generalize the GrabCut method to object segmentation from image sequence by propagating the distribution through sequence. The effectiveness and robustness of the proposed OOI extraction method for LFR image sequences is demonstrated by extensive experiments on real-world LFR sequences collected by the VISAT™ MMS.

Acknowledgments

This paper is supported by National Natural Science Foundation of China (No. 40971245).

References

1. P. L. Correia and F. Pereira, "Classification of video segmentation application scenarios," *IEEE Trans. Circuits Syst. Video Technol.* **14**(5), 735–741 (2004).
2. V. Kolmogorov et al., "Bi-layer segmentation of binocular stereo video," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, San Diego, CA, USA, pp. 407–414, IEEE (2005).
3. Y. Li, J. Sun, and H. Y. Shum, "Video object cut and paste," *Acm. T. Graphic.* **24**(3), 595–600 (2005).
4. J. Wang et al., "Interactive video cutout," *Acm. T. Graphic.* **24**(3), 585–594 (2005).
5. J. Malcolm, Y. Rathi, and A. Tannenbaum, "Multi-object tracking through clutter using graph cuts," in *Proc. International Conference on Computer Vision*, Rio de Janeiro, Brazil, pp. 2726–2730, IEEE (2007).
6. P. Yin et al., "Tree-based classifiers for bilayer video segmentation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, MN, USA pp. 295–302, IEEE (2007).
7. G. Bae, S. Kwak, and H. Byun, "Interactive object segmentation system from a video sequence," in *Proc. International Symposium on Human Interface*, San Diego, CA, USA, pp. 221–228, Springer-Verlag (2009).
8. C. Y. Chung and H. H. Chen, "Video object extraction via MRF-based contour tracking," *IEEE Trans. Circuits Syst. Video Technol.* **20**(1), 149–155 (2010).
9. Y. Z. Yu, A. Ferencz, and J. Malik, "Extracting objects from range and radiance images," *IEEE Trans. Vis. Comput. Graphics* **7**(4), 351–364 (2001).
10. N. Bourbakis, "Visual target tracking, extraction and recognition from a sequence of images using the LG graph approach," *Int. J. AI Tools* **11**(4), 513–529(2002).
11. C. H. Wang and L. Guan, "Graph Cut video object segmentation using Histogram of Oriented Gradients," in *Proc. IEEE International Symposium on Circuits and Systems*, Seattle, WA, USA, pp. 2590–2593, IEEE (2008).
12. A. Criminisi et al., "Bilayer segmentation of live video," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, New York, NY, USA, pp. 53–60, IEEE (2006).
13. A. Hernandez et al., "Spatio-Temporal GrabCut human segmentation for face and pose recovery," in *Proc. Computer Vision and Pattern Recognition Workshops*, San Francisco, CA, USA, pp. 33–40, IEEE (2010).
14. J. Ning et al., "Scale and orientation adaptive mean shift tracking," *IET Computer Vision* **6**(1), 52–61 (2012).
15. G. Petrie, "An introduction to the technology mobile mapping systems," *GeoInformatics* **13**(1), 32–43 (2010).
16. D. Li, J. Huang, and Z. Shao, "Digital earth with digital measurable images," in *ISPRS Congress*, Beijing, CRC Press/Balkema, Taylor & Francis Group (2008).
17. S. Hinz, "Automatic object extraction for change detection and GIS update," in *ISPRS Congress*, Beijing, CRC Press/Balkema, Taylor & Francis Group (2008).
18. C. Wang et al., "Automatic road vector extraction for mobile mapping systems," in *ISPRS Congress*, Beijing, CRC Press/Balkema, Taylor & Francis Group (2008).
19. A. Ruta, Y. M. Li, and X. H. Liu, "Real-time traffic sign recognition from video by class-specific discriminative features," *Pattern Recogn.* **43**(1), 416–430 (2010).
20. W. Liu et al., "A system for road sign detection, recognition and tracking based on multi-cues hybrid," in *Proc. IEEE Intelligent Vehicles Symposium*, Xian, China, pp. 562–567, IEEE (2009).
21. C. Rother, V. Kolmogorov, and A. Blake, ""GrabCut"—Interactive foreground extraction using iterated graph cuts," *Acm. T. Graphic.* **23**(3), 309–314 (2004).
22. Y. Y. Boykov and M. P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images," in *Proc. IEEE International Conference on Computer Vision*, Vancouver, British Columbia, Canada, pp. 105–112, IEEE (2001).
23. Y. Boykov and G. Funka-Lea, "Graph cuts and efficient N-D image segmentation," *Int. J. Comput. Vision* **70**(2), 109–131 (2006).
24. J. Lezama et al., "Track to the future: spatio-temporal video segmentation with long-range motion cues," in *IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, USA, IEEE (2011).
25. H. L. Li, K. N. Ngan, and Q. Liu, "FaceSeg: automatic face segmentation for real-time video," *IEEE Trans. Multimedia* **11**(1), 77–88 (2009).
26. Y. Mu et al., "Automatic video object segmentation using graph cut," in *Proc. International Conference on Image Processing*, Atlanta, GA, USA, pp. 1505–1508, IEEE (2007).
27. B. L. Price, B. S. Morse, and S. Cohen, Ieee, "LIVEcut: learning-based interactive video segmentation by evaluation of multiple propagated cues," in *Proc. IEEE International Conference on Computer Vision*, Kyoto, Japan, pp. 779–786, IEEE (2009).
28. M. J. Wu et al., "Segmenting and tracking multiple objects under occlusion using multi-label graph cut," *Comput. Electr. Eng.* **36**(5), 927–934 (2010).
29. D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(5), 603–619 (2002).
30. Z. W. Zhu et al., "Combining Kalman filtering and mean shift for real time eye tracking under active IR illumination," in *Proc. International Conference on Pattern Recogniton*, Quebec, Canada, pp. 318–321, IEEE (2002).
31. V. Karavasilis, C. Nikou, and A. Likas, "Visual tracking by adaptive Kalman filtering and mean shift," in *Hellenic Conference on Artificial Intelligence*, Athens, Greece, Springer-Verlag (2010).
32. A. Yilmaz, "Object tracking by asymmetric kernel mean shift with automatic scale and orientation selection," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, MN, USA, pp. 140–145, IEEE (2007).
33. N. S. Peng, J. Yang, and Z. Liu, "Mean shift blob tracking with kernel histogram filtering and hypothesis testing," *Pattern Recogn. Lett.* **26**(5), 605–614 (2005).

34. R. T. Collins, "Mean-shift blob-tracking through scale space," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Madison, WI, USA, pp. 234–240, IEEE (2003).
35. D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *IEEE Trans. Pattern Anal. Mach. Intell.* **25**(5), 564–577 (2003).
36. A. Yilmaz, "Kernel-based object tracking using asymmetric kernels with adaptive scale and orientation selection," *Mach. Vision. Appl.* **22**(2), 255–268 (2011).
37. K. M. Yi, H. S. Ahn, and J. Y. Choi, "Orientation and scale invariant mean shift using object mask-based kernel," in *Proc. International Conference on Pattern Recognition*, Tampa, FL, USA, pp. 3121–3124, IEEE (2008).
38. S. A. Vigus, D. R. Bull, and C. N. Canagarajah, "Video object tracking using region split and merge and a Kalman filter tracking algorithm," in *Proc. International Conference on Image Processing*, Thessaloniki, Greece, pp. 650–653, IEEE (2001).
39. X. H. Li et al., "Object tracking using an adaptive Kalman filter combined with mean shift," *Opt. Eng.* **49**(2), 020503 (2010).

Peng Li received his BE and MEsc in electrical engineering from the National University of Defense Technology in China in 2005 and 2007. He is currently pursuing his PhD in ATR National Laboratory at National University of Defense Technology and is studying in Xiamen University, China, as an exchange student. His research interests include image processing, computer vision, remote sensing, and pattern recognition.

Cheng Wang is currently a professor at the Department of Computer Science, School of Information Science and Technology, Xiamen University, China. He received the PhD degree (2002) in Information and Communication Engineering from the National University of Defense Technology, China. His research focuses on remote sensing image processing, mobile mapping data analysis, video enhancement, microscopy image analysis, and multisensor fusion. He is the co-chair of Work Group 3/I in the International Society of Photogrammetry and Remote Sensing (ISPRS), the council member of the Chinese Society of Image and Graphics (CSIG), and the member of SPIE and IEEE.